COMPUTERS AND BIOMEDICAL RESEARCH 2, 486–493 (1969)

A Mathematical Approach to Medical Diagnosis: Application to Polycythemic States Utilizing Clinical Findings with Values Continuously Distributed *

CARTER R. BISHOP AND HOMER R. WARNER

Department of Medicine and
Department of Biophysics and Bioengineering,
University of Utah,
Salt Lake City, Utah 84112

Received May 12, 1969

Certain clinical findings, largely laboratory data, which help to distinguish the various polycythemic states, are continuously distributed. These continuous distributions were used as probability functions from which elements of Bayes' formula for conditional probability were derived. This formula was written into a computer program which calculates a probability value for each of the diagnostic possibilities. A preliminary trial of the program employing this technique was applied to 103 cases that were either normal or known to have polycythemia rubra vera. When presented with the initial patient data the mathematical probability diagnosis was correct 95% of the time. Three hematologists were 76% correct whereas three general practitioners were correct in 65% of the same cases.

In the process of making a medical diagnosis the physician must recall a great deal of past information against which the new patient is compared. This recollection is always subject to human error. Even with the best recollection, the physician would be unlikely to compare each case equally and he might be influenced by findings which have little bearing on the diagnosis. The computer has been employed to provide perfect recollection and impose internal consistency.

Several workers have proposed conditional probability as a means of accomplishing these goals using Bayes' formula for conditional probability.[1,2,3,4,5] These programs have dealt with clinical findings that could be classified in a given patient as being present or absent. When clinical findings which present as a continuous function were utilized, probability values were assigned to certain ranges of the continuous function and often such data have been reduced to a binary function by setting a dividing line, one side of which was considered

* This investigation was supported by a postdoctoral research fellowship (1-F2-GM-29,729) from the National Institute of General Medical Sciences and by grants from the Division of Research Facilities and Resources (FR-00012) and the National Institute of Arthritis and Metabolic Diseases (AM-04489), National Institutes of Health, Bethesda, Maryland.

TABLE 1

LIST OF CLINICAL FINDINGS EVALUATED FROM THE HISTORY, PHYSICAL EXAM AND THE LABORATORY

Volume of Packed Red Cells
Red Blood Cell Count
Hemoglobin Concentration
White Blood Cell Count
Granulocyte Count
Platelet Count
Leukocyte Alkaline Phosphatase
Age at the Onset of the Disease
Basophilia
Splenomegaly
Ph^1 Chromosome
Sex
Race
Altitude of Patient's Place of Residence

abnormal[3,4,6]. It is the purpose of this paper to present an approach to such clinical findings which treats the continuous distribution as a probability function and thus permits the data to speak more sensitively for themselves.

The polycythemic states were divided into five disease categories; normal, polycythemia rubra vera [PRV], secondary polycythemia (A) [2° (A)], secondary polycythemia (B) [2° (B)], and chronic myelocytic leukemia [CML]. For the purposes of this preliminary report only normal and PRV are considered. Thus the probability space is made up of patients referred to a hematologist because of polycythemia who were proved to be normal or have PRV. The records of 250 cases of polycythemia were examined at the Division of Experimental Medicine, University of Oregon College of Medicine, Portland, Oregon.*

The clinical findings recorded at the patients' initial visits to the University of Oregon Medical Center were used. If prior therapy might have altered any of the findings, no recording of that value was made. Clinical findings used in this study are listed in Table 1. For those findings that were binary in character (e.g. Ph^1-chromosome, splenomegaly, sex and basophilia), the incidence of that finding in the two disease categories was calculated. A histogram of the frequency distribution was prepared in each of the disease categories for each of the eight clinical findings that were continuously distributed.

The frequency distribution of the laboratory findings which were continuous functions were well described by a lagged-normal distribution.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left[\frac{x-\mu}{\sigma}\right]^2} - \tau \cdot f'(x), \tag{1}$$

* The records were made available by Dr. E. E. Osgood and his associates.

where $f'(x)$ is d $f(x)/dx$. An essential property of this equation is that the variance of $f(x)$ is a function of σ and τ, so that S.D.$^2 = \sigma^2 + \tau^2$. The best fitting lagged-normal distribution [7] was found for each histogram (Fig. 1). From the mean, standard deviation, and tau parameter of each of these distributions the probability of any given value for a particular test in each disease category can be determined.

Table 2 shows the mean (μ), standard deviation (σ) and tau (τ) values for these eight tests in the two disease categories. In the normal group, each of the variables was normally distributed, so no entry is made for τ which is zero.

In Table 3 the incidence of the Ph^1 chromosome, basophilia, spelenomegaly and male sex are depicted. The probability for the Ph^1 chromosome in the normal and polycythemia rubra vera patients is assumed to be near 0.0, although this figure was not derived from the 250 records reviewed. Basophilia is only very roughly measured here as the presence of any basophils in a count of 200 cells. Splenomegaly is considered present if the spleen is felt below the costal margin and absent if not palpable.

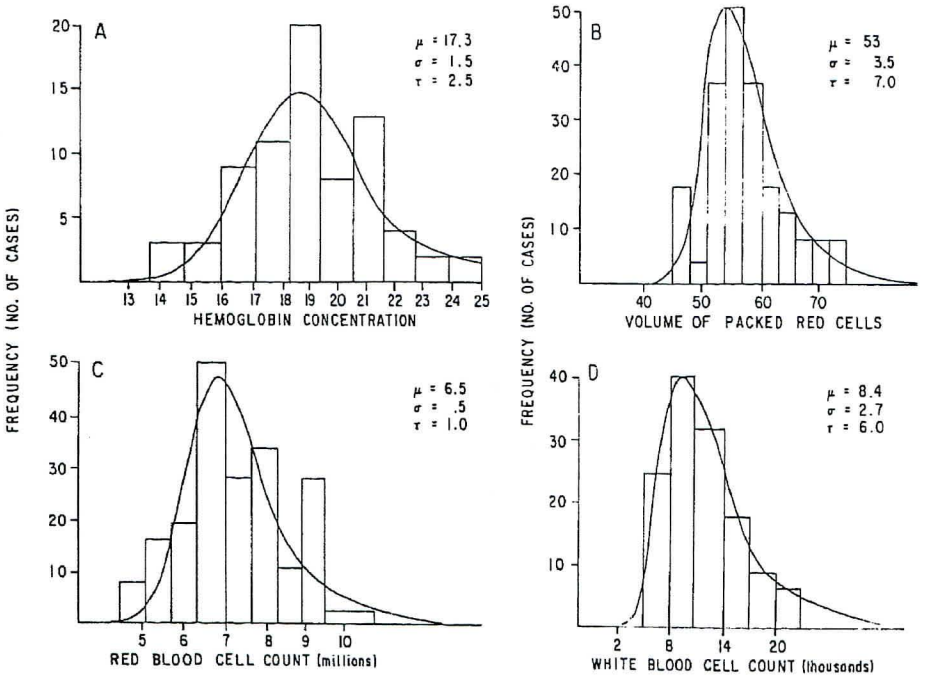

FIG. 1. Histogram showing the frequency of various blood determinations in polycythemia Rubra Vera with the Best Fitting Lagged-Normal Distribution. The bar graph is a simple frequency plot of the blood determination. The smooth curve is the best fitting lagged-normal distribution.

TABLE 2

MEAN (μ), STANDARD DEVIATION (σ) AND SKEW FACTOR (τ) OF THE CLINICAL FINDINGS IN THE DISEASE CATEGORIES; NORMAL AND POLYCYTHEMIA RUBRA VERA *PRV*

Clinical findings	Normal[c]		*PRV*		
	μ	σ	μ	σ	τ
VPRC[a]	48.3	3.19	53.0	3.5	7.0
Red cell count	5.25	.52	6.5	.5	1.0
Hemoglobin conc.	16.6	1.60	17.3	1.5	2.5
White cell count	8.0	2.0	8.4	2.7	6.0
Granulocyte count	6.27	8.2	5.5	1.5	5.0
Platelet count	244.6	60.3	100.0	40.0	190.0
Age at onset	47.3	12.1	56.6	12.1	0.0
LAP[b]	44.5	12.5	233.4	77.3	0.0

[a] *VPRC*—Volume of packed red cells. [b] *LAP*—Leukocyte alkaline phosphatase; τ—0.0 leaves a normal distribution. [c] τ—0.0 for all distributions in Normal subjects.

TABLE 3

PROBABILITY OF BINARY FINDINGS IN TWO DISEASE CATEGORIES

	Norm	*PRV*
Ph[1] chromosome	.0001	.0001
Basophilia	.34	.52
Splenomegaly	.001	.66
Male sex	.5	.575

Bayes' formula for conditional probability is as follows:

$$P_{(Di \mid Sj)} = \frac{P_{(Di)} \cdot \prod_{j=1}^{n} P_{(Sj \mid Di)}}{\sum_{i=1}^{m} P_{(Di)} \cdot \prod_{j=1}^{n} P_{(Sj \mid Di)}}, \tag{2}$$

where D_i is one of a set of 'm' diseases which are mutually exclusive, S_j is one of a set of 'n' clinical findings. The $P_{(D_i)}$ is the *a priori* probability of the disease in the population under consideration. It is this form of the equation that is used when the eight clinical findings that are continuously distributed are being considered. For those factors which were binary, the equation takes the following form (2):

$$P_{(Di \mid Sj)} = \frac{P_{(Di)} \prod_{j=1}^{n} \{P_{(Sj \mid Di)} \cdot a_j + (1 - P_{(Sj \mid Di)})(1 - a_j)\}}{\sum_{i=1}^{m} [P_{(Di)} \prod_{j=1}^{n} \{P_{(Sj \mid Di)} \cdot a_j + (1 + P_{(Sj \mid Di)})(1 - a_j)\}]}. \tag{3}$$

The 'a' is a factor which is assigned a value of 1.0 if the clinical finding is prestnt and 0.0 if the finding is absent. This term can be used to weight the importance of some of clinical signs as was done in this program with the basophil count. Since basophilia is only grossly quantitative, its presence or absence as defined should not carry the weight that the volume of packed red cells does. For that reason 'a' is 0.6 if basophilia were present and 0.4 if absent.

Figure 2 shows the distribution of hemoglobin values in the two disease categories. If a hypothetical case had a hemoglobin concentration of 18.0 gm %, the probability that a hemoglobin value would be in that range in each of the five disease categories ($P_{(S/D)}$ in Eq. (2) can be determined by integrating each distribution from 18.0 gm % minus .05 standard deviation to 18 gm % plus .05 standard deviation. The reason for this should be apparent from Fig. 2, since 18.0 gm % is closest to the mean value for PRV (18.9) but the height of the line is greatest under the normal curve (mean 16.6). Therefore if $P_{(S/D)}$ were estimated by integration of each curve over the same range, Normal would be assigned a higher probability than PRV. By integrating over a range which is a function of σ and τ, this problem is avoided and the proper values are obtained.

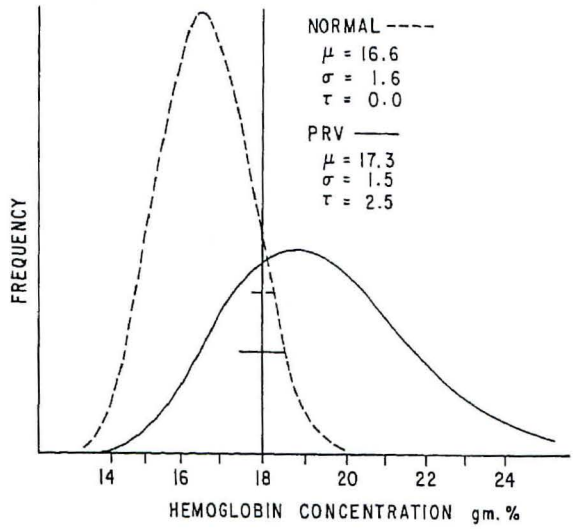

FIG. 2. Frequency distribution of hemoglobin concentration values in the two disease categories. The normal distribution which describes the data from known normal cases is plotted together with the lagged-normal distribution describing the data from the PRV group. A vertical line at 18.0 gm % represents a value from an undiagnosed case. The horizontal lines are proportional to the standard deviations of the respective curves. It is over a range proportional to these lines that the curves are integrated to obtain the probability of the hemoglobin concentration of 18.0 gm % given the disease.

PRELIMINARY TRIAL

One hundred and three cases were selected at random from the known cases from the University of Oregon Medical Center which were not the same cases used to make up the data matrix. By follow-up examination, these patients were normal or had polycythemia rubra vera but all were initially suspected of having polycythemia. The clinical findings from these patients were examined by three hematologists and three general practitioners and were evaluated by the program as well. For the purpose of this trial, the priori probability of normal was set at 0.8 and of PRV 0.2, in order to minimize the chance of a false positive diagnosis of PRV. The doctors were clearly informed that the cases were either normal or PRV. The results of this trial can be seen on Table 4. All of the polycythemia vera cases missed by the computer were also misdiagnosed by three or more of the six physicians. The two cases misdiagnosed as having polycythemia rubra vera by the computer were also misdiagnosed by four of the six physicians. The average hematological values in the cases misdiagnosed by the doctors as normal, were within one standard deviation of the mean values for patients with polycythemia rubra vera.

It is of note that three of the symptoms used are interdependent to varying degrees: volume of packed red cells, red blood cell count and hemoglobin concentration. To assess the effect this has on the ability of the program to make a

TABLE 4

COMPARISON OF PERFORMANCE OF PROGRAM WITH HEMATOLOGISTS AND GENERAL PRACTITIONERS ON 103 CASES WITH POLYCYTHEMIA RUBRA VERA OR NORMAL

	Correct	False positive	False negative
Program	95	2	3
Hematologists	76	2	22
General practitioners	65	1	34

TABLE 5

COMPARISON OF PROGRAM PERFORMANCE WHEN SELECTED MEASUREMENTS OF RED BLOOD CELLS ARE USED ALONG WITH NON-RED CELL PARAMETERS

Measurements used	Correct	False +	False −
VPRC RBC Hgb	95	2	3
RBC Hgb	93	4	3
RBC VPRC	93	5	2
VPRC Hgb	90	6	4
Hgb	89	3	8

correct diagnosis, the three combinations of two of the red cell parameters were used to the exclusion of the third in three trial runs. Also, since the hemoglobin concentration was the determination most regularly recorded in these test cases, it was used to the exclusion of VPRC and RBC in an additional trial run. The results of these trials can be seen in Table 5. With the combination of two of the three determinations used in conjunction with the non-red cell parameters, there was little or no significant difference in the results. When the hemoglobin concentration alone was used along with the non-red cell parameters, the results were not quite as good.

DISCUSSION

Immediate therapy is not crucial to the survival of the patient with polycythemia rubra vera. However, it is essential that a correct diagnosis be established with certainty before instituting therapy with radioactive materials. Therefore, it is best to have the probability of false positive diagnoses at a minimum particularly if potentially dangerous therapy such as radiophosphate or cytotoxic drugs are to be used. For this purpose the *a priori* probability figure can be adjusted favoring the diagnosis of normal.

As experience with this program is gained in a given clinic, the matrix data can be updated to reflect this experience. Periodically, a reevaluation of the various clinical findings can be made to eliminate from consideration in diagnosing future cases those variables which do not help in the differential diagnosis. For instance, it is anticipated that WBC will be redundant as all cases will have the more specific granulocyte count GBC. If this proves true, WBC can be eliminated. Additional findings can be evaluated as well, and added to the program if they are found to contribute to making the correct diagnosis. It should be apparent that such laboratory data as pulmonary function studies, arterial gas determinations, erythropoietin determinations and x-ray findings could be included in such a prospective evaluation. Some of these determinations are readily available at many medical centers and as experience with these findings is gained, they can be added to the program.

The frequency distribution of the data for the symptoms in the polycythemia vera group is best described by a lagged-normal distribution. In the other disease categories a normal distribution often described the data best. As long as the function describing the data is normalized to unit area it makes no difference what function is used. However, it is important that the function chosen to represent the distribution of the data describes the data as accurately as possible. In this study the critical part of separation is at the lower limits of the PRV group and the upper limit of the normal group and the use of a lagged-normal to describe the PRV data gave sharper separation than could be obtained by approximating the data with a normal distribution.

Vanderplas [8] has pointed out that interdependence of symptom subsets should

be tested lest a spurious posterior probability be assigned a given disease. Though we know that the volume of packed red cells, red blood cell count and hemoglobin concentration are interdependent, the effect of deleting one or two of these parameters does not significantly alter the overall performance of the program though it may change the performance in an isolated case.

The improved accuracy of diagnosis made possible by the computer in the case of the hematological disorder presented here (95% compared to 65 and 76%) is even more striking than was reported for congenital heart disease.[2,9,10] Perhaps this is due to the difficulty the physician has in interpreting data which are continuously distributed (such as a WBC) compared to binary (yes-no) data (such as the presence or absence of a murmur). The method described here offers a general approach to the problem of diagnosis since both binary and continuously distributed data may be utilized effectively together.

Since modern technology has greatly improved the accuracy of laboratory determinations and the rapidity with which they can be obtained, the development of new methods for analysis of laboratory data and clinical findings becomes imperative. The approach to medical diagnosis described above is a step toward helping the physician extract more accurate information from a patient's data which will be of value in making decisions about that patient.

REFERENCES

1. Nugent, C. A., Warner, H. R., Dunn, J. T., and Tyler, F. H. Probability theory in diagnosis of Cushing's Syndrome. *J. Clin. Endocrinol. Metab.* **24,** 621 (1964).
2. Warner, H. R., Toronto, A. F., Veasy, L. G., and Stephenson, R. A mathematical approach to medical diagnosis: application to congenial heart disease. *J. Am. Med. Assoc.* **177,** 177 (1961).
3. Overall, J. E. and Williams, C. M. Conditional probability program for diagnosis of thyroid function. *J. Am. Med. Assoc.* **183,** 307 (1963).
4. Lodwick, G., Haun, C. L. Smith, W. E., Keller, R. F., and Robertson, E. D. Computer diagnosis of primary bone tumors, a preliminary report. *Radiology* **80,** 273 (1963).
5. Henschke, U. K. and Flehinger, B. J. Decision theory in cancer therapy. *Cancer* **20,** 1819 (1967).
6. Fitzgerald, L. T., Clyde, M. W., and Overall, J. E. A computer program for diagnosis of thyroid disease. *Am. J. Roentgenol.* **97,** 901 (1966).
7. Nicholes, K. K., Warner, H. R., and Wood, E. H. A study of dispersion of an indicator in the circulation, *Ann, N.Y. Acad. Sci.* **115,** 721 (1964).
8. Vanderplas, J. M. A method for determining probabilities for correct use of Bayes' theorem in medical diagnosis. *Computers and Biomed. Res.* **1,** 215 (1967).
9. Toronto, A. F., Veasy, G., and Warner, H. R. Evaluation of a computer program for diagnosis of congenital heart disease. *Progr. Cardiovascular Diseases* **5,** 362 (1963).
10. Warner, H. R., Toronto, A. F., and Veasy, G. Experience with Bayes' theorem for computer diagnosis of congential heart disease. *Ann. N.Y. Acad. Sci.* **115,** 558 (1964).